DOI 10.1515/cclm-2013-1038 — Clin Chem Lab Med 2014; 52(7): 965–972

Dietmar Stöckl, Katleen Van Uytfanghe, Stefan Van Aelst and Linda M. Thienpont*

A statistical basis for harmonization of thyroid stimulating hormone immunoassays using a robust factor analysis model

Abstract

Background: Between-method equivalence ideally is achieved by calibration against an SI-traceable reference measurement procedure. For measurement of thyroid stimulating hormone (TSH), it is unlikely to accomplish this goal in mid-term. Therefore, we investigated a statistical alternative based on a factor analysis (FA) model.

Methods: The FA model was applied to TSH results for 94 samples generated by 14 immunoassays (concentration range: 0.0005–78 mIU/L). The dataset did not fulfill the assumption of a homogeneous sample from an elliptically symmetric distribution, and, therefore, required standardization prior to application of the FA model. As outliers and missing values also occurred, the key quantities of the FA model had to be estimated with a method that can handle these complications. We selected a robust alternating regressions (RAR) method, which replaces in the minimization criterion of the fitting process the squared differences between results x_{ij} and model fit $\hat{x}_{ij}$ by a weighted absolute difference. The weights are adaptively determined in successive regressions, which down weighs the outliers. The weights for missing values are set to zero.

Results: The quality of the estimated targets was reflected by their central position in the distributions, and description of the relationship between results and targets by a simple two-parameter regression equation with high correlation coefficients and low SDs of the percentage-residuals. Mathematical recalibration eliminated the method differences and improved the between-method CV from 11% to 6%.

Conclusions: RAR applied to a multimethod comparison dataset hampered by outliers and missing values, is fit to the purpose of harmonization.

Keywords: factor analysis model; harmonization; principal component analysis; robust alternating regressions.

***Corresponding author: Linda M. Thienpont,** Laboratory for Analytical Chemistry, Faculty of Pharmaceutical Sciences, Ghent University, Harelbekestraat 72, 9000 Gent, Belgium, Phone: +32 9 2648104, Fax: +32 9 2648198, E-mail: linda.thienpont@ugent.be

Dietmar Stöckl: STT-consulting, Horebeke, Belgium
Katleen Van Uytfanghe: Laboratory for Analytical Chemistry, Faculty of Pharmaceutical Sciences, Ghent University, Gent, Belgium
Stefan Van Aelst: Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Gent, Belgium

Introduction

In 2010 the AACC started an initiative to prominently promote harmonization of measurands for which no generally accepted reference measurement procedures (RMPs) exist or are under development [1, 2]. A special Task Force was mandated with proposing a generic technical harmonization process. One proposal came down to applying a new measurement standard, i.e., a panel of native sera assigned with targets statistically-derived from a multimethod intercomparison study [3]. The key to the success of this approach lies of course in the use of an adequate statistical model. It should assume that each method in the comparison tries to measure the same target quantity, which may be the true or a hypothetical reference value. The quantity is measured with both random and systematic errors. While random errors reflect the method imprecision, the systematic one, also called the bias, consists of a fixed, method-specific and a random, sample-specific part. The latter is due to method non-specificity, and, hence, depends on the concentration of interfering substances. The method-specific bias reflects to what extent a method consistently over/underestimates the target quantity. Its magnitude is usually not constant, but depends on the amount of quantity present in the sample. In the absence of a RMP to set trueness-based targets for a split-sample method comparison, the bias components cannot be identified/quantified, nor eliminated, but the statistical model can estimate so-called composite reference values from the data that can subsequently be used to set each method's results to a common scale [4, 5]. Statistically speaking, these values are the realizations of an unobservable random variable, called a factor. Henceforth, the model is called factor analysis (FA) model involving one

factor. Principal component analysis (PCA) is the standard approach to estimate the key quantities of the FA model, i.e., composite reference values as the (suitably centered) scores of the first principal component. However, PCA estimates are only reliable if the method comparison data are a homogeneous sample that does not contain any outliers, missing values, heterogeneity or skewness.

Recalibration against consensus values estimated by a FA model to bring multiple sets of results into closer mutual alignment has been exemplary described in the past, but never knew a real breakthrough [4–6]. Therefore, we recently investigated the potential in an application to derive the all-procedure trimmed mean from method comparison data by several insulin immunoassays [7]. The insulin dataset was of particular interest, because it had been used before to prove the feasibility of trueness-based standardization by adopting a candidate RMP [8]. The conclusion was that recalibration against statistically-derived targets resulted in an equivalent quality compared to the RMP approach.

Being involved in an IFCC effort towards standardization of thyroid stimulating hormone (TSH) assays, we took the opportunity to investigate the statistical approach to harmonization. Indeed, to the best of our knowledge, no SI-traceable RMP is available yet and/or likely to be on mid- to long-term. From a previous method comparison study we knew that most of the immunoassays had their calibration set-point within 10% from the overall mean (apart from a few showing discrepancies of approx. 30%–40%), with an excellent correlation and low dispersion of regression residuals [9]. We found these observations sufficiently convincing to challenge the technical feasibility of aligning TSH immunoassays to target concentrations set by an FA model applied on measurement data from a recent method comparison. We opted for robust alternating regressions (RAR) instead of PCA and emphasize its validity to handle our data [10–13]. We verified in particular whether recalibration to the estimated target values adequately removed method-specific biases, so that the remaining dispersion nearly entirely could be attributed to within-method effects.

Materials and methods

Data collection

TSH data were obtained from a method comparison with 14 immunoassays and 94 clinical samples (concentration range 0.0005–78 mIU/L). Their identity of the assays is not disclosed, because this is of no relevance for the statistical study (for a more extended rationale, see the Data Supplement that accompanies the article at http://www.degruyter.com/view/j/cclm.2014.52.issue-7/issue-files/cclm.2014.52.issue-7.xml). The samples were analyzed in duplicate in the same run. Seven samples with concentrations below the most common functional sensitivity (<0.012 mIU/L) and for which some immunoassays had not reported results, were included for statistical estimation of the target values, but excluded from validation and recalibration.

Descriptive statistics

The effect of random errors was reduced by working on the mean of the replicates. The data were collected in a 94 (rows representing the different samples)×14 (columns for the different methods) matrix. The matrix contained 40 missing values (3%).

Measurement heterogeneity and other complications (skewed data distributions, outliers) were investigated from presenting the data for the individual methods/samples in histograms, box- and barplots. In addition, for each method and sample the mean, median and SD were calculated as well as the median absolute deviation (MAD) given by $mad(x_i)=med(|x_i-med(x_i)|)$.

Statistical approach

We applied the FA model to statistically estimate composite reference values for the samples. Since our dataset was heterogeneous and contained outliers, we first robustly standardized the data by using the median and MAD. The target values estimated by the FA model were later mapped back to the original data scale. With x_{ij} being the measurement result of the jth method for the ith sample, the FA model containing one factor can be written as $x_{ij}=\alpha_j+\beta_j f_i+\varepsilon_{ij}$, where the values f_i are the composite reference values for the samples. In statistical terms, the values f_i are unobservable realizations of a latent random variable F. In the FA model it is assumed that the method-specific bias of a method j consists of a constant and proportional part referred as α_j and β_j, respectively. The constant bias is the systematic deviation between a method's measurement and the target quantity, and is usually rather small. In contrast, the proportional bias is the systematic deviation that depends on the amount of target quantity. The error term ε_{ij} contains both the imprecision and random bias of the method.

We used an alternating regression method that directly estimates the unknown parameters in the FA model by minimizing the sum of the deviations between the measured values x_{ij} and their corresponding model fit $\hat{x}_{ij}=\hat{\alpha}_j+\hat{\beta}_j\hat{f}_i$. To handle missing values, the method excludes the corresponding residual in the minimization criterion. If the deviations are measured by squared differences, i.e., minimize $\sum_{i,j}(x_{ij}-\hat{x}_{ij})^2$, then the alternating least squares regressions technique is obtained [10, 11]. In the robust variant (RAR), the squared difference is replaced by a weighted absolute difference, i.e., minimized $\sum_{i,j}w_{ij}|x_{ij}-\hat{x}_{ij}|$. The weights w_{ij} are determined adaptively by the fitting procedure, which down weighs the outliers in the successive regression fits [12]. Missing values are still handled by exclusion in the minimization criterion, i.e., by setting the weight w_{ij} equal to zero.

Assessment of quality

We assessed the validity of the RAR procedure from a statistical point of view by making dotplots of the data for each sample by all methods, with inclusion of the estimated composite reference values. We also examined which fraction of the heterogeneity in the measurements can be explained by the FA model. This fraction can be measured by the robust R-squared, given by $RR^2 = 1 - \left(\frac{\sum_{i,j} w_{ij} |x_{ij} - \hat{x}_{ij}|}{\sum_{i,j} w_{ij} |x_{ij} - \hat{\alpha}_j|} \right)^2$.

In the absence of proportional biases, RR^2 is reduced to 0. However, in the opposite case, RR^2 differs from zero and the FA model can explain a substantial fraction of the heterogeneity. We also presented the measurement results of each method in a scatter plot versus the RAR estimated targets.

To assess the quality of the RAR-derived target concentrations from an analytical point of view, we compared for each of the methods the magnitude of the percentage-difference for the individual samples from their respective target against total error limits derived from biological variation (19.1%) [14]. Second, we did correlation and regression analysis of the results by the immunoassays and the estimated target concentrations. Finally, we calculated the percentage-residuals for each of the samples, by deriving the 'predicted mean of the duplicate' based on the best fit between the mean and the respective target values (using weighted linear regression or power functions). The residuals were expressed in percent relative to the 'predicted mean of the duplicate'. We also calculated for each method the SD of the percentage-residuals.

Recalibration

We used the RAR-derived target concentrations for mathematical recalibration of the data by each individual method. We applied either weighted linear regression or power functions depending on the best fit of the data. We verified the quality of recalibration again in percentage-difference plots with inclusion of the aforementioned total error limits [14]. Finally, we evaluated the pre- and post-recalibration between-method variation for the individual samples ('between-method CV'). The CV was calculated from the ratio between the SD on the results per sample by all methods to the mean. Statistical analysis and plotting were done with the freely available advanced statistical software R (R core team 2012) version 2.15.1 [15], Microsoft Excel® 2010 and CBstat5 (K. Linnet, Denmark).

Results

Descriptive statistics

The boxplots in Figure 1 show that for all methods the distribution of the measurement results was strongly skewed, and that the spread differed between the methods (compare, e.g., the boxplot for methods C and E with that for method I). The skewness of the distributions is also reflected by the mean concentration for the different methods ranging from 5.3 mIU/L (method I) to 8.1 mIU/L (method G), while the range of the median concentrations was from 1.5 mIU/L (method I) to 2.1 mIU/L (method K). The skewness created an even larger discrepancy between SDs (ranging from 11.2 mIU/L to 17.0 mIU/L) and MADs (ranging from 1.9 mIU/L to 2.5 mIU/L).

The discrepancies between the different methods are shown in more detail in the barplots of four particular methods (Supplemental Data, Figure S1). For method M, the data are generally close to the median, while those of method I are consistently below, with a difference proportional to the median concentration. The data of method K exceed the median, in particular, for samples with a higher concentration. The barplot for method G is

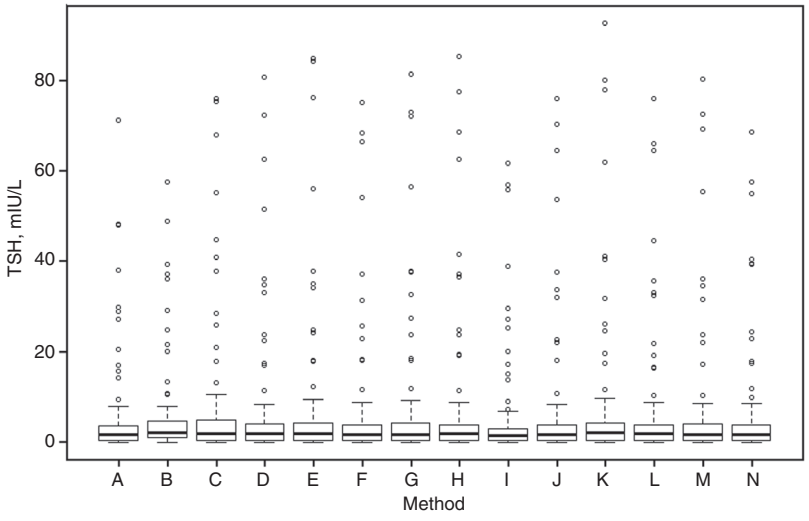

Figure 1 Boxplot showing the distribution of the measurement results per method.

an example of a dataset with missing values (17 empty bars in total).

The distribution of the results per sample is shown in four boxplots (Figure 2). The samples are ranked from the lowest to the highest median concentration. These boxplots clearly show heteroscedasticity of the spread in results between the methods. Moreover, they reveal that many samples contain outliers. Outliers were identified by using the standard rule associated with boxplots, i.e., results that lie further than one and a half times the interquartile range below or above the box are considered outliers. The data heterogeneity is confirmed in the barplots in Supplemental Data, Figure S2. They show the distribution of the measurements by all methods (A–N) for two individual samples. While most of the methods had their results close to the median, some deviated considerably (either higher or lower). Note that the empty bar in the right plot means that method G did not report a result for that sample.

Statistical approach

The histograms and corresponding estimated density curves in Figure 3 illustrate that, after robust data standardization, there is still a lot of heterogeneity in the distributions of the methods. For example, the histograms for methods A, B, H, I still clearly are asymmetric, while methods C, E and L contain outlying data. Due to these complications, we used iterative RAR to estimate the composite reference values. The procedure converged quickly (after only 9 iterations) and the statistically-derived concentrations ranged from 0.0005 mIU/L to 78.4 mIU/L.

Assessment of quality

The dotplots in Supplemental Figure S3 show for each sample the measurement data by all methods and the estimated composite reference value. They illustrate that

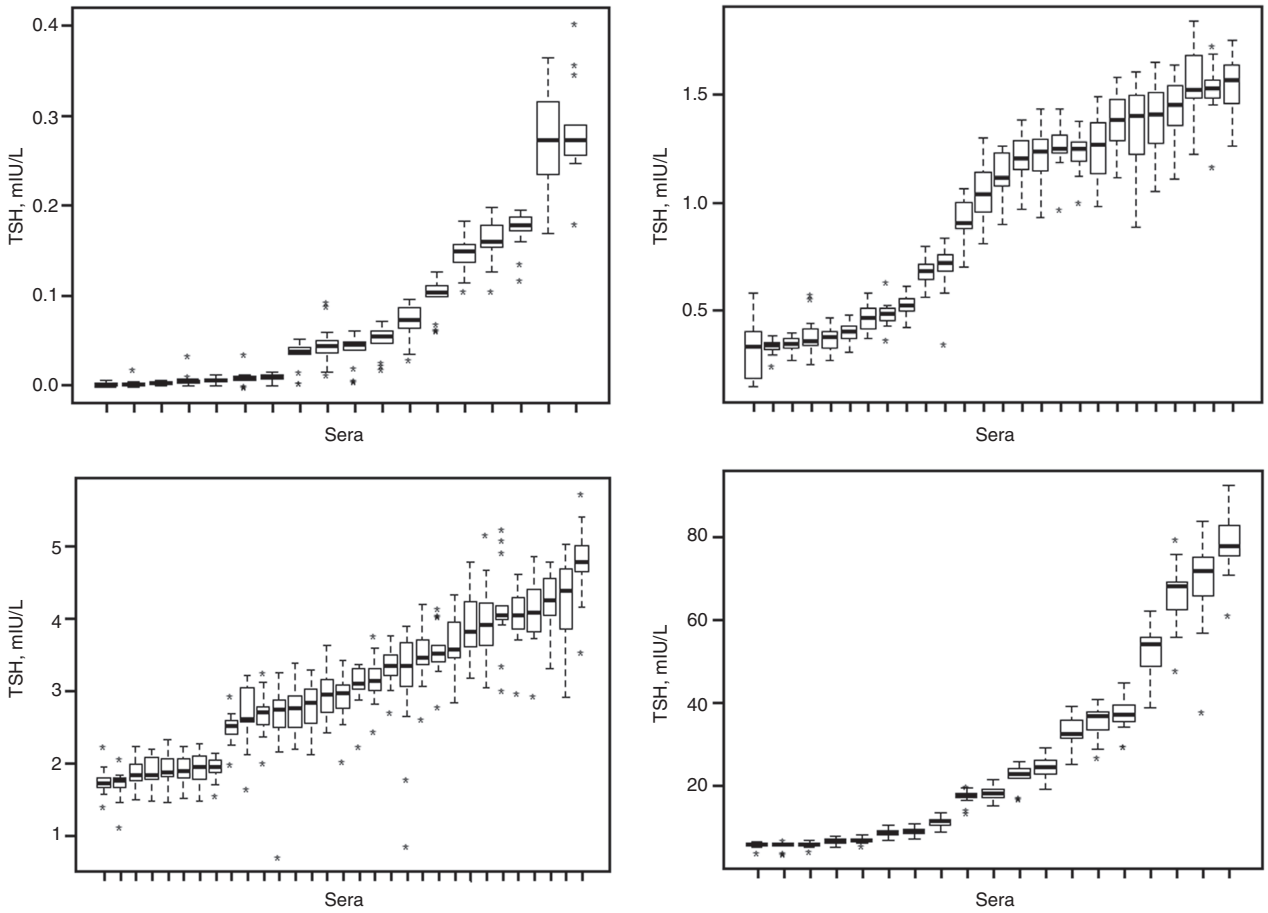

Figure 2 Boxplots showing the distribution of results per sample.
Note the four different TSH concentration ranges in the ordinate (0.00–0.3 mIU/L; 0.3–1.6 mIU/L; 1.6–5 mIU/L and 5–80 mIU/L). * indicates the outliers. Samples are arranged in ascending order of the median concentrations estimated from the measurement data by the 14 methods.

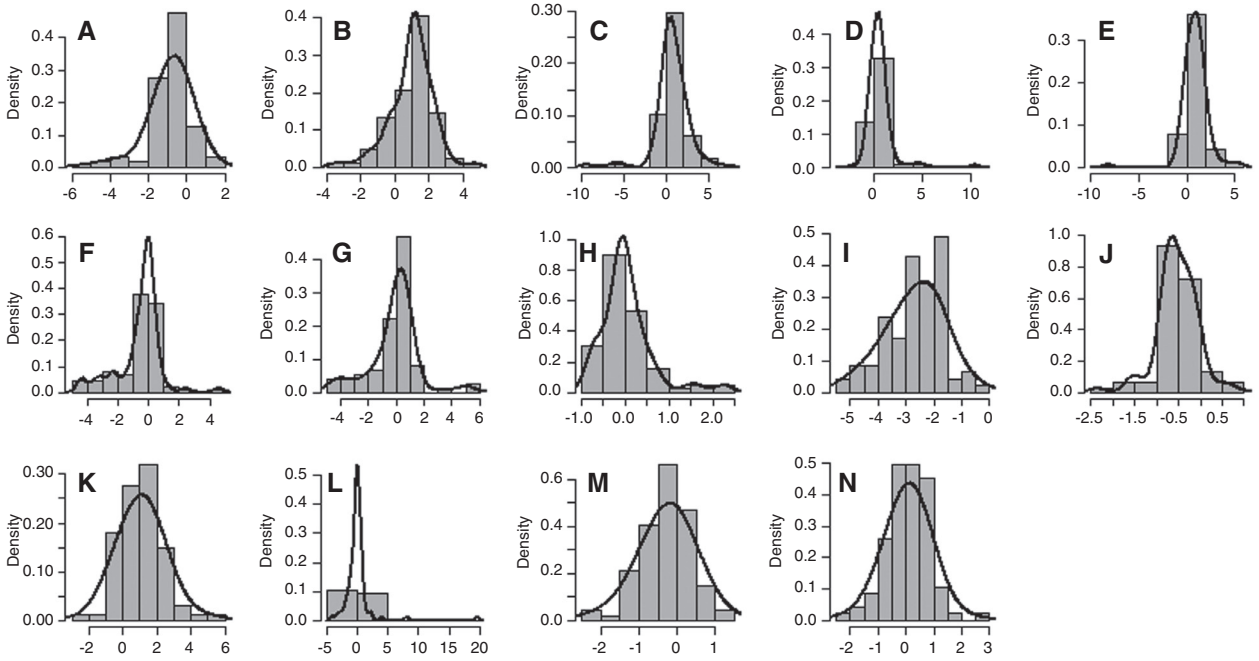

Figure 3 Histograms with corresponding estimated density curves of the distribution for each of the methods after robust standardization of the measurement data.

the latter lies for each sample in the center of the data, without being affected by the outlying measurements or the heterogeneity in distributions. In addition, a robust R-squared of 0.183 was obtained, which means that 18.3% of the variability between the methods can be explained by proportional deviations from the estimated composite reference values.

Supplemental Figure S4 gives for all methods the constant and proportional part of the method-specific bias estimated by RAR. The scatterplots in Supplemental Figure S5 show how well the bias components have been estimated by RAR, e.g., for method H, for which the RAR procedure estimated a large proportional bias with a constant bias at zero, the scatterplot confirms that most data lie above the first bisector and that the distance to the line proportionally increases. The scatterplot for method M confirms the small biases estimated by RAR. Only for method I one sees a large proportional bias in the scatterplot (all data for samples with a high concentration are below the bisector, thus strongly deviate from the estimated targets), while estimated by RAR it was almost zero.

Figure 4A represents the percentage-difference of the results by all immunoassays from the statistically estimated target concentrations before recalibration. Plots for the individual methods can be found in Supplemental Figure S6. Table 1 lists the correlation coefficients, the SDs of the percentage-residuals, and the applied regression functions. These plots and table show the analytical quality of the estimated composite reference values as a basis for recalibration. Of all data, 86% were within the ±19.1% biological total error limits, while the data for one method (K) were completely out; 95% of the data were within ±25% limits. The most discrepant percentage-differences were situated in the low concentration range. For the majority of the immunoassays weighted linear regression analysis could be applied, while for only three immunoassays, a power function was used. The range for r was from 0.9946 to 0.9996, for the SDs of the percentage-residuals from 3.4 to 9.7, with the exception of one outlying SD of 18.3 (immunoassay G). Note that this SD was highly influenced by the percentage-residuals of the two samples with the lowest concentration (≤0.044 mIU/L). After omission of these samples from the dataset, the SD decreased to 7.8%.

Recalibration

Figure 4B represents the percentage-difference plot for all immunoassays after mathematical recalibration (for the individual plots, see Supplemental Figure S6). All immunoassays had the majority of the differences (98% of all data) within the biological total error limit. Note in addition that even 95% were within ±14%. Mathematical

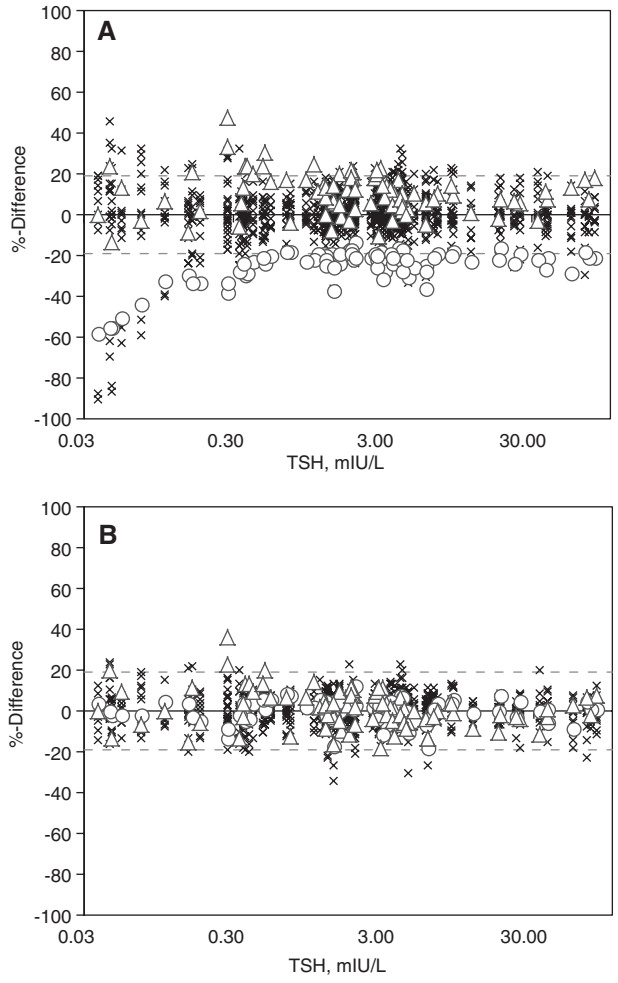

Figure 4 Percent-difference plot of the measurement results (mean of duplicates) by all immunoassays from the statistically-estimated target values.
(A) Before; (B) after mathematical recalibration. Note the logarithmic scale of the abscis. Triangles: method K with the highest deviation before recalibration; circles: method I with the lowest deviation before recalibration; dashed line: ±19.1% total error limit based on biological variation.

recalibration resulted in a considerable decrease in between-method CV (Figure 5). For samples with a concentration <0.15 mIU/L, it decreased in average from 30% to approximately 8%, while in the higher concentration range from 11% to approximately 6%.

Discussion

The TSH data used in this setting were from a multimethod comparison study. The purpose was to investigate whether consensus values obtained by data treatment with a reliable FA model could serve as new measurement

Table 1 Correlation coefficients, SDs of the percentage-residuals, and type of regression function applied.

Method	Correlation coefficient	$SD_{\%\text{-residuals}}$	Regression function
J	0.9993	3.4	WLR[a]
H	0.9996	4.4	Power
D	0.9995	5.2	Power
I	0.9982	5.7	WLR
A	0.9993	5.9	Power
M	0.9976	6.3	WLR
N	0.9969	7.3	WLR
E	0.9961	7.9	WLR
B	0.9963	8.1	WLR
C	0.9956	8.7	WLR
L	0.9946	9.5	WLR
K	0.9948	9.5	WLR
F	0.9980	9.7	WLR
G	0.9981	18.3	WLR

[a]Weighted linear regression.

standard, so that the multiple sets of measurement results come in closer alignment. In the FA model, it is assumed that the method-specific biases consist of a constant and proportional part. Although the assumption of a linear relationship between the methods may not be satisfied completely, it often is a good first order approximation with sufficient accuracy to reliably estimate consensus values [4]. Our recalibration data confirmed this.

We first considered PCA, which is the standard estimation procedure for the FA model. It obtains the estimates by decomposition of the covariance matrix of the data, implying the assumption of multivariate data being a homogeneous sample from an elliptically symmetric distribution. We scrutinized our dataset for satisfaction of this assumption, which was essential to obtaining unbiased estimates. We started with descriptive statistical data

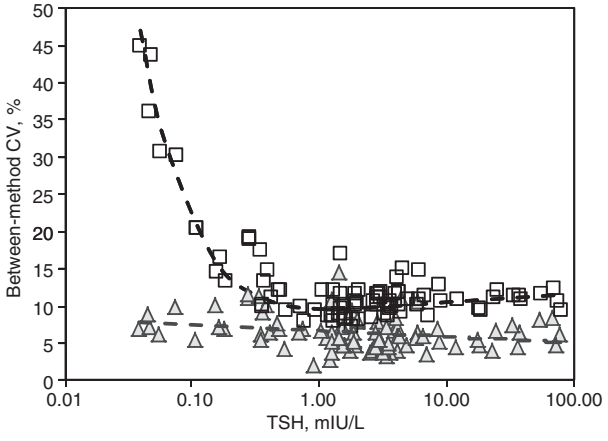

Figure 5 Between-method CV observed before and after mathematical recalibration.

analysis, and found huge heterogeneity. We considered transformation techniques to remove the skewness or outliers, i.e., logarithmic and more flexible Box-Cox ones, however, none were successful. Moreover, such transformations have a large impact on the relations between the methods, which may complicate successive data analysis. Instead, we performed robust data standardization, which reduced the skewness in the distributions of all methods, but not to the required extent (see the remaining non-symmetric histograms with outliers in Figure 3). To reduce the effect of outliers, we could have identified and trimmed them with a robust PCA method [7, 13]. However, this approach requires the existence of a large enough homogeneous subsample in the data from which the covariance matrix can be estimated. Moreover, still, the robust PCA approach is not able to solve the issue of missing values. Removal of these samples from data treatment was not an option, because it causes loss of information, which in turn may bias the outcome. Therefore, we finally decided to apply a robust variant of an alternating regressions method, i.e., RAR. This direct estimation approach does not rely on the covariance matrix of the data and can easily handle outliers. In contrast to PCA methods, it effectively handles missing values by setting their weight w_{ij} equal to zero instead of removing an entire row (sample) or column (method) from the data matrix. We could have considered alternative models, e.g., a mixed effects model, i.e., with estimation of a fixed effect for each sample and a random effect for each method. The former fulfills the role of the consensus value, and allows that each assay has the same constant deviation (the latter effect) from this value, irrespective of the sample [16–18]. However, the heterogeneity, skewness and outliers in our data were prohibitive to using this alternative, because they do not satisfy the assumptions made by standard fitting methods for mixed models, i.e., that due to random error effects, the data follow a normal distribution with constant variance. Moreover, methods for transformation of the data to remove the heterogeneity and/or skewness without affecting the validity of the mixed model, are, if not impossible, hard to find. Also other assumptions about the relations between the methods were not satisfied, i.e., that for each sample the variations among the methods (after transformation) behave according to some fixed regular distribution (usually a normal distribution). Although this assumption can be relaxed by including an interaction term, it may not be sufficient to capture all heterogeneity between the samples. To the best of our knowledge, estimation methods for the mixed model that can effectively handle the presence of skewness and outliers are not directly available.

For all samples, the composite reference values obtained by RAR were lying well in the center of the distributions. This proves their quality and validity for recalibration. The added value of the FA model to explain data heterogeneity is confirmed by the fact that nearly 1/5 of the between-method heterogeneity could be explained by the systematic biases. Thus, not surprisingly, almost all methods showed a considerable systematic deviation component, apart from method N. The constant and proportional bias part was for most methods well estimated, except for method I. Although this apparently conflicts with the required robustness of the RAR procedure, it is explained by the fact that RAR heavily down weighs all data with large deviations from the mean in the high concentration range, which of course prevents estimation of the large proportional bias. Notwithstanding this, the composite reference values were reliably estimated, while the scatterplot reveals the large proportional bias afterwards.

Whether recalibration of the methods on the basis of the statistical consensus concentrations will result in equivalence of results depends not only on the quality of the estimates, but also on the influence of the methods' random bias in the process. In this regard, it is important that the matrix-specific bias components are maximally down weighted, which requires inclusion of as many well-correlating methods as possible in the intercomparison. For heterogeneous analytes, this necessitates investigation whether all methods measure the same measurand by performing cluster analysis [6]. This requirement was fulfilled in our study with 14 immunoassays with previously proven high correlation to the overall mean [9]. We confirmed by several means good quality of the estimated recalibration basis. First the relationship of the results of all immunoassays with the targets could be described by a simple two-parameter regression analysis (Table 1). In addition the scatter around the regression line was good to excellent for the majority of the immunoassays, so were the correlation coefficients (min. 0.9946). As a result, mathematical recalibration was straightforward and eliminated the concentration-dependent differences. This was even better reflected by the improvement of the between-method CV. In other words, we can claim that future implementation of a panel of samples bearing RAR-estimated target concentrations as new measurement standard for harmonization of TSH immunoassays, looks promising.

Notwithstanding the success of this study, some cautionary notes and limitations need to be stressed. First, the input data always will determine the quality of the statistical output, thus the success of future, similar

studies will depend on interplay of several factors. One is the study design requiring use of a high-end panel of samples and use of an experimental measurement design optimized towards obtaining input data of high quality (e.g., replicate measurements of all samples in the same run to reduce both random error and calibration effects). Also a high correlation between the methods and preferably homogeneous data distribution are extremely important for a successful outcome, because they warrant use of a simple best-fitting statistical model and estimates with low uncertainty. In case of heterogeneous data complicated by skewed distributions, outliers and/or missing values, adequate statistical solutions exist, however, they need to be validated for the quality of the estimates. Last but not least, statistically-derived target concentrations are not necessarily trueness-linked. Therefore, before starting any harmonization activity, it is highly recommendable to carefully investigate whether there are no standardization efforts in the pipeline.

Acknowledgments: The authors are indebted to the IFCCs Committee for Standardization of Thyroid Functions Test and the partners from the in-vitro diagnostic industry for the inspirational support to conduct this statistical study.

Conflict of interest statement

Authors' conflict of interest disclosure: The authors stated that there are no conflicts of interest regarding the publication of this article. Research support played no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.
Research funding: None declared.
Employment or leadership: None declared.
Honorarium: None declared.

Received December 2, 2013; accepted January 27, 2014; previously published online February 22, 2014

References

1. Miller WG, Myers GL, Gantzer ML, Kahn SE, Schönbrunner ER, Thienpont LM, et al. Roadmap for harmonization of clinical laboratory measurement procedures. Clin Chem 2011;57:1108–17.
2. Joint Committee for Traceability in Laboratory Medicine (JCTLM) Database of higher-order reference materials, measurement methods/procedures and services. Available from: http://www.bipm.org/jctlm/. Accessed 31 July, 2013.
3. Thienpont LM, Van Houcke SK. Traceability to a common standard for protein measurements by immunoassay for in-vitro diagnostic purposes. Clin Chim Acta 2010;411:2058–61.
4. Lawton WH, Sylvestre EA, Young-Ferraro BJ. Statistical comparison of multiple analytic procedures: application to clinical chemistry. Technometrics 1979;21:397–409.
5. Carey RN, Wold S, Westgard JO. Principal component analysis: an alternative to "referee" methods in method comparison studies. Anal Chem 1975;47:1824–9.
6. Rymer JC, Sabatier R, Daver A, Bourleaud J, Assicot M, Bremond J, et al. A new approach for clinical biological assay comparison and standardization: application of principal component analysis to a multicenter study of twenty-one carcinoembryonic antigen immunoassay kits. Clin Chem 1999;45:869–81.
7. Van Houcke SK, Van Aelst S, Van Uytfanghe K, Thienpont LM. Harmonization of immunoassays to the all-procedure trimmed mean – proof of concept by use of data from the insulin standardization project. Clin Chem Lab Med 2013;51:e103–5.
8. Miller WG, Thienpont LM, Van Uytfanghe K, Clark PM, Lindstedt P, Nilsson G, et al. Toward standardization of insulin immunoassays. Clin Chem 2009;55:1011–8.
9. Thienpont LM, Van Uytfanghe K, Beastall G, Faix JD, Ieiri T, Miller WG, et al. Report of the IFCC Working Group for Standardization of Thyroid Function Tests, Part 1: Thyroid-stimulating hormone. Clin Chem 2010;56:902–11.
10. Wold H. Nonlinear estimation by iterative least squares procedures. In: David FN, editor. Research papers in statistics: Festschrift for Jerzy Neyman. London: Wiley, 1966;411–44.
11. Gabriel KR, Zamir S. Lower rank approximation of matrices by least squares with any choice of weights. Technometrics 1979;21:489–98.
12. Croux C, Filzmoser P, Pison G, Rousseeuw PJ. Fitting multiplicative models by robust alternating regressions. Statist Comput 2003;13:23–36.
13. Hubert M, Rousseeuw PJ, Van Aelst S. High-breakdown robust multivariate methods. Stat Sci 2008;23:92–119.
14. Westgard QC. Desirable specifications for total error, imprecision, and bias derived from biologic variation. Available from: http://www.westgard.com/biodatabase1.htm. Accessed 31 July, 2013.
15. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org/. Accessed 31 July, 2013.
16. Chu H, Chen S, Louis TA. Random effects models in a meta-analysis of the accuracy of two diagnostic tests without a gold standard. J Am Stat Assoc 2009;104:512–23.
17. Joo DJ, Jung I, Kim MS, Huh KH, Kim H, Choi JS, et al. Comparison of the affinity column-mediated immunoassay and microparticle enzyme immunoassay methods as a tacrolimus concentration assay in the early period after liver transplantation. Transplan Proc 2010;42:4137–40.
18. Rose CE, Romero-Steiner S, Burton RL, Carlone GM, Goldblatt D, Nahm MH, et al. Multilaboratory comparison of Streptococcus pneumoniae opsonophagocytic killing assays and their level of agreement for the determination of functional antibody activity in human reference sera. Clin Vaccine Immunol 2011;18:135–42.